SPEQTACLE: An automated generalized fuzzy C-means algorithm for tumor delineation in PET

Jérôme Lapuyade-Lahorgue, Dimitris Visvikis, Olivier Pradier, Catherine Cheze Le Rest, Mathieu Hatt

SPEQTACLE: an automated generalized fuzzy C-means algorithm for tumor delineation in PET

Jérôme Lapuyade-Lahorgue[1], Dimitris Visvikis[1], Olivier Pradier[1,2], Catherine Cheze Le Rest[3], Mathieu Hatt[1]

[1]LaTIM, INSERM, UMR 1101, Brest, France

[2] Radiotherapy department, CHRU Morvan, Brest, France

[3]DACTIM, Nuclear medicine department, CHU Milétrie, Poitiers, France

Corresponding author: M. Hatt

INSERM, UMR 1101,LaTIM

CHRU Morvan, 2 avenue Foch

29609 Cedex, Brest, France

Tel: +33(0)2.98.01.81.11

Fax: +33(0)2.98.01.81.24

E-mail: hatt@univ-brest.fr

Wordcount: ~9400 (8900 for main body, 540 for appendix, 255 for abstract)

Disclosure of Conflicts of Interest: No potential conflicts of interest were disclosed.

Funding: No specific funding.

Abstract

Purpose: accurate tumordelineation in PET images is crucial in oncology. Although recent methods achieved good results, there is still room for improvement regarding tumors with complex shapes, low signal-to-noise ratio and high levels of uptake heterogeneity.

Methods: We developed and evaluated an original clustering-based method called SPEQTACLE (Spatial Positron Emission Quantification of Tumor - AutomatiCLp-norm Estimation), based onthe fuzzy C-means (FCM)algorithm with a generalization exploiting a Hilbertian norm to more accurately account for the fuzzy and non-Gaussian distributions of PET images.An automatic and reproducibleestimation scheme of the norm on an image-by-image basis wasdeveloped.Robustness was assessed by studying the consistency of results obtained on multiple acquisitions of the NEMA phantom on three different scanners with varying acquisitions parameters. Accuracy was evaluatedusing classification errors (CE) onsimulated and clinical images. SPEQTACLE was compared to another FCM implementation (FLICM)and FLAB.

Results: SPEQTACLE demonstrateda level of robustness similar to FLAB (variability of 14±9% *vs.* 14±7%, p=0.15) and higher than FLICM (45±18%, p<0.0001), and improved accuracy with lower CE(14±11%) over bothFLICM (29±29%) and FLAB (22±20%) on simulated images. Improvement was significant for the more challenging cases with CE of 17±11% for SPEQTACLE*vs.* 28±22% for FLAB (p=0.009) and 40±35% for FLICM (p<0.0001). For the clinical cases, SPEQTACLE outperformed FLAB and FLICM (15±6% *vs.* 37±14% and 30±17%, p<0.004).

Conclusions: SPEQTACLE benefitted from the fully automatic estimation of the norm on a case-by-case basis.This promising approach will be extended to multimodal imagesand multi-class estimation in future developments.

Keywords: PET segmentation - clustering methods -Fuzzy C-means-Hilbertian norm.

Introduction

Positron Emission Tomography (PET) is established as a powerful tool in numerous oncology applications[1], including target definition in radiotherapy planning[2], and therapy monitoring[3, 4], two applications for which tumor delineation is an important step, allowing for instance further quantification of PET images such as the extraction of image based biomarkers[5–7]. Within this context, automatic 3D functional volume delineation presents a number of advantages relative to manual delineation which is tedious, time-consuming and suffers from low reproducibility[8].PET imaging is characterized by lower spatial resolution (~4-5mm 3D full width at half maximum (FWHM)) and signal-to-noise ratio (SNR) compared to other medical imaging techniques such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT). In addition, the existing large variability in scanner models and associated reconstruction algorithms (and their parameterization) leads to PET images with varying properties of textured noise, contrast, resolution and definition in clinical routine practice, which becomes acritical issue in multi-centric clinical trials[9]. Thus, automatic, repeatable and accurate, but also robustsegmentation of tumor volumes is still challenging.Many methods based on various image segmentation paradigms, including but not limited to fixed and adaptive thresholding, active contours and deformable models, region growing, statistical and Markovian models, watershed transform and gradient, textural features classification, and fuzzy clustering,have been already proposed[10, 11]. Despite the recent improvements and the high level of accuracy and robustness achieved by some of these state-of-the art methods, there is still room for improvement, especially regarding the delineation of tumors with complex shapes, high level of uptake heterogeneity, and/or low imageSNR.

Methods including clustering and Bayesian estimation have demonstrated promising performance in PET tumor volume segmentation.

On the one hand, in Bayesian segmentation methods, statistical distributions (also called noise distributions)of the intensities are modeled by summarizing the histogram of the

images considering a reduced number of parameters to estimate. These methods provide automatic algorithms allowing noise modeling and *prior* solution selection,which allows them in turn to be less sensitive to noise than other segmentation approaches due to their statistical modeling[12].Bayesian segmentation methods can be viewed as regularized “blind” statistical approachesin which the prior probability constraints the solution. This prior distribution can be defined in different ways according to the targeted application, for instance usinghidden Markov field or chain models where the *prior* distribution is a Markov field distribution[13]. Relatively recent examples of such methods specifically developed for PET include Fuzzy Hidden Markov Chains (FHMC) and the Fuzzy Locally Adaptive Bayesian (FLAB) methods. In FHMC, the prior distribution was modeled using fuzzy hidden Markov chains[14], whereas in FLAB the 3D neighborhood of a given voxel was used to locally estimate the fuzzy measurefor each voxel [15, 16], leading to a more accurate segmentation of small structures. FLAB can be considered to be one of the state-of-the-art methods for PET, according to its wide success due to its robustness, its repeatability and its overall accuracy demonstrated on both simulated and various clinical datasets including radiotracers of hypoxia and cellular proliferation [8, 16–23].

On the other hand, clustering methods aim at partitioning the images into clusters depending on the statistical properties of the voxel intensities. The main interest of clustering methods compared to Bayesian methods lies in their low computational cost, as well as easier parameters estimation and overall implementation. The most known and used clustering method is the K-means clustering which has been extended to Fuzzy C-means clustering (FCM) by considering a fuzzy instead of a deterministic measure on the cluster’s membership.The Fuzzy C-means (FCM) algorithm has several advantages including flexibility and low computational cost. However, it fails to correctly address non-Gaussian noise, geometrical differences between clusters, spatial dependency between voxels, as well as the variability of fuzziness and noise properties or textures of the PET images that arise

from the large range of PET image reconstruction algorithms and post-reconstruction filtering schemes currently used in clinical practice.

Regarding FCM more specifically, amongst the other different generalizations of FCM, some incorporate a more accurate description of the clusters' geometry in the data model, for example by replacing the Euclidian norm by the Mahalanobis distance[24]. This method requires estimating the covariance matrices of each cluster additionally to the centers of the clusters and therefore takes into account that the clusters are not necessarily of identical sizes. Another version uses the Lebesgue l^1 and l^∞ norms instead of the Euclidian norm[25]. Other authors have proposed to replace this Euclidian norm by a Hilbertiankernel[26], which is more reliable in cases where the data does not follow a Gaussian mixture model. Finally, other authors have replaced the probability measure by evidential measure as in the "possibilistic" FCM[27]. This last approach is interesting within the context of evidential theory, however the way hard decision is carried out is heuristic and difficult to justify[28]. Amongst the methods exploiting the spatial information, it was proposed to generalize FCM by introducing spatial constraints to regularize it[29]. Other methods, such as the Fuzzy Local Information C-Means (FLICM)algorithm, incorporate in the minimization criteria the distance between voxels[30].

The goal of this work was to focus on FCM and to propose a novel generalization in order to improve on the accuracy without sacrificing on robustness of PET tumor segmentation results compared to current state-of-the-art techniques, for challenging heterogeneous tumours.We have chosen to generalize FCM using a Hilbertian kernel,with the norm parameter not set empirically or *a priori*but rather estimated on an image-by-image basis, using a fully automatic scheme based on a likelihood maximization algorithm. The new algorithm was compared to FLICM and FLAB in terms of robustness and accuracy on real and simulated PET image datasets.

Materials and methods

A. FCM algorithm and its extensions

Classical FCM algorithm

The FCM algorithm consists in finding for each class $i \in \{1,\cdots,C\}$, where C is the number of classes, and for each voxel $u \in V$ of the finite set of voxels $V \subset \Re^3$, the centers $\mu_i \in \Re$ and the degrees of belief $p_{u,i} \in [0,1]$ minimizing the criterion:

$$\sum_{u \in V} \sum_{i=1}^{C} p_{u,i}^{m} \left| y_u - \mu_i \right|^2 \qquad (1)$$

under the constraint: $\sum_{i=1}^{C} p_{u,i}^{m} = 1$,

where y_u is the observed intensity for the voxel u and the parameter $m > 1$ controls the fuzzy behavior and is usually chosen as $m = 2$.

Thedetails regarding this minimization are provided in appendix A.

Regarding the segmentation, for each voxel $u \in V$, the class $i \in \{1,\cdots,C\}$ maximizing the probability $p_{u,i}$ is chosen. This decision step is the same for the generalized FCM (GFCM).

FCM as a Bayesian inference method

The traditional “hard” K-means clustering is equivalent to a Bayesian method where the observations are modeled as a Gaussian mixture. FCM clustering can also be rewritten in order to highlight a *prior* distribution regarding the parameters $p_{u,i}$ and μ_i, and a likelihood associating observations with the parameters. This idea has already been exploited by choosing *prior* distributions to optimize the estimation[31]. The minimization of eq. (1) is equivalent to the maximization of:

136 $P = f\left(\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u - \mu_i\right|^2\right)$, where f is a positive function such that P is a probability

137 density according to the observed variables $(y_u)_{u\in V}$ called “likelihood”. From statistics, the

138 maximization of P is equivalent to a likelihood maximization and is exhaustive (*i.e.* uses the

139 entire information of the sample) if the density of $(y_u)_{u\in V}$ maximizes the Shannon entropy.

140 Moreover, one can show that a distribution whose form is given by

141 $P = f\left(\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u - \mu_i\right|^2\right)$ is an elliptical distribution (*i.e.*isodensities are ellipsoid) with

142 center $\frac{\sum_{i=1}^{C} p_{u,i}^{m}\mu_i}{\sum_{i=1}^{C} p_{u,i}^{m}}$ and dispersion given by $\frac{1}{\sqrt{\sum_{i=1}^{C} p_{u,i}^{m}}}$. An elliptical distribution is entirely

143 determined by its functional parameter f, its center and its dispersion. Amongst the elliptical

144 distributions with the same center and dispersion, one can show that the maximum entropy is

145 reached if f is an exponential function.

146 Consequently, in this case, the minimization of eq. (1) is equivalent to the maximization of:

147
$$\begin{aligned} P &= \exp\left(-\frac{1}{2}\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u - \mu_i\right|^2\right) \\ &= \prod_{u\in V}\exp\left(-\frac{1}{2}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u - \mu_i\right|^2\right) \end{aligned} \qquad (2)$$

148 Also:

149
$$\begin{aligned} &\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u - \mu_i\right|^2 \\ &= \left(\sum_{i=1}^{C} p_{u,i}^{m}\right)\times\left(y_u^2 - 2\frac{\sum_{i=1}^{C} p_{u,i}^{m}\mu_i y_u}{\sum_{i=1}^{C} p_{u,i}^{m}} + \frac{\sum_{i=1}^{C} p_{u,i}^{m}\mu_i^2}{\sum_{i=1}^{C} p_{u,i}^{m}}\right). \end{aligned}$$

Consequently, conditionally to the parameters, the observations y_u are independent and Gaussian distributed as:

$$p(y_u|(\mu_i)_{1\le i\le C},(p_{u,i})_{1\le i\le C}) := N\left(\frac{\sum_{i=1}^{C} p_{u,i}^m \mu_i}{\sum_{i=1}^{C} p_{u,i}^m}, \frac{1}{\sqrt{\sum_{i=1}^{C} p_{u,i}^m}}\right) \quad (3)$$

Whereas the *prior* distribution for parameters is given by:

$$p((\mu_i)_{1\le i\le C},(p_{u,i})_{1\le i\le C,u\in V}) \propto \prod_{u\in V} \frac{1}{\sqrt{\sum_{i=1}^{C} p_{u,i}^m}} \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{C} p_{u,i}^m \mu_i^2 - \frac{\left(\sum_{i=1}^{C} p_{u,i}^m \mu_i\right)^2}{\sum_{i=1}^{C} p_{u,i}^m}\right)\right]$$

Drawbacks of the classical FCM

The previous theory results in two major drawbacks:

(a). FCM clustering is equivalent to a maximum *posterior* estimation when the observations follow a Gaussian distribution conditionally to the parameters. Consequently, FCM leads to inaccurate estimation when the data are not Gaussian.

(b). Similarly, FCM clustering assumes that the observations are independent conditionally to the parameters, leading to inaccurate segmentation in the presence of spatial dependencies.

B. SPEQTACLE algorithm: an automatic Generalized FCM algorithm (GFCM)

In this work we investigated the advantage of generalizing FCM by considering the Hilbertian l^p-norm instead of the Euclidian norm and providing an associated scheme that enables a fully automated estimation of the norm parameter for optimal delineation on a case-by-case basis, in order to reduce user interaction and avoid empirical optimization. Indeed, a user-

168 defined choice of the norm parameter based on visual analysis seems challenging because
169 of its non-intuitive nature, and would suffer from low reproducibility. An alternative would be
170 to optimize empirically the norm value on a training dataset, although it is unlikely that a
171 single norm value would be appropriate for all cases. We have consequently developed an
172 approach to automatically estimate the norm value for each image.

173 The proposed algorithm is called Spatial Positron Emission Quantification of Tumor
174 volume:AutomatiCL^p-norm Estimation (SPEQTACLE).

175 *Principle of GFCM algorithm*

176 In the GFCM algorithm, the minimization criterion becomes:

177 $$\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u-\mu_i\right|^{\alpha} \qquad (4)$$

178 where, the norm parameter $\alpha>1$ and with no solution for $\alpha=1$. Moreover, the cluster centers
179 μ_i cannot be estimated explicitly when $\alpha\neq 2$, whereas $\alpha=2$ corresponds to the standard
180 FCM. When $\alpha>2$ and $\alpha<2$, the centers are computed using the Newton-Raphson
181 algorithm and gradient descent respectively (for details we refer the reader to Appendix B.
182 and C.).

183 *Generalized Gaussian distribution*

184 We assume that conditionally on the parameters $(\mu_i)_{1\le i\le C}$ and $(p_{u,i})_{1\le i\le C}$ the observation is
185 approximately distributed as a generalized Gaussian distribution whose density is

$y \to \frac{\alpha}{2\sigma\Gamma\left(\frac{1}{\alpha}\right)} \exp\left(-\frac{|y-\mu|^{\alpha}}{\sigma^{\alpha}}\right)$ parameterized by a center $\mu = \frac{\sum_{i=1}^{C} p_{u,i}^{m} \mu_i}{\sum_{i=1}^{C} p_{u,i}^{m}}$, a dispersion $\sigma = \frac{1}{\left(\sum_{i=1}^{C} p_{u,i}^{m}\right)^{\frac{1}{\alpha}}}$ and a shape α.

Estimation of the norm

The estimation technique presented in the next section is based on the above generalized Gaussian distribution. Contrary to the Gaussian case, it is only an approximation; indeed the expression (4) can be expressed as a product of $\sum_{i=1}^{C} p_{u,i}^{m}$ and a term of form $|y_u - \beta|^{\alpha}$ only in the Gaussian case, which corresponds to $\alpha = 2$. However, it becomes a generalized Gaussian distribution if $p_{u,i} = 1$ holds for only one class. This approximation is valid as long as the probabilities $(p_{u,i})$ are not too far from the configuration $p_{u,i} = 1$. Consequently, the norm parameter has to be estimated from an area for which one can consider that $p_{u,i} = 1$ holds. In practice, this area was automatically selected using a background subtraction method in order to provide a first guess of the tumor region, as recently proposed[32]. In order to simplify the estimation task, we have chosen to estimate the norm for this background-subtracted region, which is likely to correspond to a first estimation of the tumor region, and set a single norm parameter value for all classes.

The next step involves the estimation of the different parameters using likelihood maximization.

Let us denote $\mu = \frac{\sum_{i=1}^{C} p_{u,i}^{m} \mu_i}{\sum_{i=1}^{C} p_{u,i}^{m}}$ and $\sigma = \frac{1}{\left(\sum_{i=1}^{C} p_{u,i}^{m}\right)^{\frac{1}{\alpha}}}$.

First, one can assume that these values do not depend on u and secondly, that the distribution of the observations y_u in the selected area is approximately the generalized Gaussian distribution. Let $(y_u)_{u\in W}$ be the sample from the selected area W, the maximum likelihood estimators of μ, σ and α, denoted $\hat{\mu}_{ML}$, $\hat{\sigma}_{ML}$ and $\hat{\alpha}_{ML}$ are solutions of the system:

a. $\sum_{u\in W} \mathrm{sgn}(y_u - \hat{\mu}_{ML}) \left| y_u - \hat{\mu}_{ML} \right|^{\hat{\alpha}_{ML}-1} = 0$;

b. $\hat{\sigma}_{ML}^{\hat{\alpha}_{ML}} = \hat{\alpha}_{ML} \times \frac{1}{|W|} \sum_{u\in W} \left| y_u - \hat{\mu}_{ML} \right|^{\hat{\alpha}_{ML}}$;

c. $\frac{\hat{\alpha}_{ML} + \psi\left(\frac{1}{\hat{\alpha}_{ML}}\right)}{\hat{\alpha}_{ML}^{\,2}} = \frac{1}{|W|} \sum_{u\in W} \log\left(\frac{\left| y_u - \hat{\mu}_{ML} \right|}{\hat{\sigma}_{ML}}\right) \frac{\left| y_u - \hat{\mu}_{ML} \right|^{\hat{\alpha}_{ML}}}{\hat{\sigma}_{ML}^{\hat{\alpha}_{ML}}}$,

where, $|W|$ is the cardinality of W and ψ is the log-derivative of the Eulerian function (see AppendixD).

These equations are not linear and cannot be solved independently. Consequently, the solution is estimated by using a combination of a variational method and the Newton-Raphson algorithm as outlined below:

1. Let $\mu^{(0)}$, $\sigma^{(0)}$ and $\alpha^{(0)}$ be the initial values ;
2. From $\alpha^{(p)}$, compute $\mu^{(p+1)}$ by solving $\sum_{u\in W} \mathrm{sgn}(y_u - \mu^{(p+1)}) \left| y_u - \mu^{(p+1)} \right|^{\alpha^{(p)}} = 0$ using the Newton-Raphson algorithm;
3. From $\alpha^{(p)}$ and $\mu^{(p+1)}$, compute $\sigma^{(p+1)} = \alpha^{(p)} \times \frac{1}{|W|} \sum_{u\in W} \left| y_u - \mu^{(p+1)} \right|^{\alpha^{(p)}}$;

4. From $\mu^{(p+1)}$ and $\sigma^{(p+1)}$, compute $\alpha^{(p+1)}$ by solving

$$\frac{\alpha^{(p+1)}+\psi\left(\frac{1}{\alpha^{(p+1)}}\right)}{\alpha^{(p+1)}} = \frac{1}{|W|}\sum_{u\in W}\log\left(\frac{\left|y_u-\mu^{(p+1)}\right|}{\sigma^{(p+1)}}\right)\frac{\left|y_u-\mu^{(p+1)}\right|^{\alpha^{(p+1)}}}{(\sigma^{(p+1)})^{\alpha^{(p+1)}}}$$

5. Repeat steps 2, 3 and 4 until convergence.

Although such generalized Gaussian distributions have properties that allow for convergence of the maximum likelihood estimation, the stopping criteria has to be defined. One could assume that the estimation can be stopped when the successive values of $\alpha^{(p)}$ (resp. $\mu^{(p)}$ and $\sigma^{(p)}$) are sufficiently close to each other, using the absolute distances as stopping criteria. However, the values of the parameters can be close, whereas the distance between the resulting distributions may be large. Indeed, the smaller σ is, the more sensitive to the value of μ is the resulting density. To overcome this drawback, we used a more appropriate distance; namely the distance between distributions rather than the distance between parameters' values. This distance is defined from the Fisher information matrix(Appendix E).

It has been previously shown that the set of given parameterized distributions is a Riemannian manifold whose metric tensor is given by the Fisher information matrix[33]. More precisely, let $\Lambda=\{y\to p(y|\theta):\theta\in\Theta\}$ be a smooth manifold of statistical distributions parameterized by an open set $\Theta\subset\Re^k$, the distance between "close" distributions $y\to p(y|\theta)$ and $y\to p(y|\theta+d\theta)$ is given by:

$dl=\sqrt{(d\theta)^* I(\theta)d\theta}$, where $I(\theta)$ is the Fisher information matrix and $(d\theta)^*$ is the transpose of the vector $d\theta$.

For the generalized Gaussian random variables that we use in SPEQTACLE, the Fisher information relative to the position parameter μ, the dispersion parameter σ and the norm parameter α are given respectively by:

$$I(\mu) = \alpha(\alpha-1)\frac{\Gamma\left(\frac{\alpha-1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}$$

$$I(\alpha) = 1+\frac{1}{\alpha^2}$$

$$I(\sigma) = \frac{\alpha}{\sigma^2} \text{ and } + \frac{2(1+\alpha)}{\alpha^3}\psi\left(\frac{1}{\alpha}\right) + \frac{1}{\alpha^3}\psi\left(\frac{1}{\alpha}\right)^2 ,$$
$$+ \frac{1+\alpha}{\alpha^4}\psi'\left(\frac{1}{\alpha}\right)$$

where, Γ and ψ are the Eulerian function and its log-derivative respectively. In the norm estimation algorithm, we evaluate the distance between distributions twice; namely when $\sigma^{(p)}$ and $\alpha^{(p)}$ are recomputed. It maybe also possible to evaluate the distance when $\mu^{(p)}$ is recomputed.However, if the two other parameter sequences $\sigma^{(p)}$ and $\alpha^{(p)}$ do not vary, one can reliably assume that the parameter sequence $\mu^{(p)}$ does not vary either. As the Fisher information relative to σ is given by $I(\sigma) = \frac{\alpha}{\sigma^2}$, for fixed values of α and μ the infinitesimal distance between two generalized Gaussian distributions $p(y|\alpha,\mu,\sigma)$ and $p(y|\alpha,\mu,\sigma+d\sigma)$ is $\frac{\sqrt{\alpha}}{\sigma}d\sigma$ and the distance between $p(y|\alpha,\mu,\sigma^{(p)})$ and $p(y|\alpha,\mu,\sigma^{(p+1)})$ is given by:

$$D(\sigma^{(p)},\sigma^{(p+1)}) = \sqrt{\alpha}\left|\int_{\sigma^{(p)}}^{\sigma^{(p+1)}}\frac{d\sigma}{\sigma}\right|$$
$$= \sqrt{\alpha}\left|\log\left(\frac{\sigma^{(p+1)}}{\sigma^{(p)}}\right)\right|$$

Regarding the parameter α, the integration of the Fisher metric is not explicit and requires time consuming numerical methods. We have used the Kullback-information "metric" instead, as a good approximation of the Fisher metric when the consecutive values of $\alpha^{(p)}$ are close (Appendix E). When μ and σ are set, the Kullback information from $p(y|\alpha^{(p)},\mu,\sigma)$ to $p(y|\alpha^{(p+1)},\mu,\sigma)$ is given by:

$$K(\alpha^{(p+1)}:\alpha^{(p)}) = \log\left(\frac{\alpha^{(p+1)}\Gamma\left(\frac{1}{\alpha^{(p)}}\right)}{\alpha^{(p)}\Gamma\left(\frac{1}{\alpha^{(p+1)}}\right)}\right)+\frac{\Gamma\left(\frac{1+\alpha^{(p)}}{\alpha^{(p+1)}}\right)}{\Gamma\left(\frac{1}{\alpha^{(p+1)}}\right)}-\frac{1}{\alpha^{(p+1)}}\cdot$$

Finally, in the maximum likelihood estimation algorithm, $D(\sigma^{(p)},\sigma^{(p+1)})$ and $K(\alpha^{(p+1)}:\alpha^{(p)})$ are evaluated when the value of $\sigma^{(p+1)}$ and $\alpha^{(p+1)}$ are respectively computed. The stopping rule is a fixed threshold value ε =10^{-7}small enough to ensure convergence.

C. Algorithm evaluation methodology

Repeatability and dependency of the norm estimation on initial tumor region

In order to evaluate the repeatability of SPEQTACLE, the whole process (background-subtracted area definition used to estimate the norm, followed by the iterative estimation of the norm and the modified FCM clustering) was applied 20 times to the same tumorimages.

In order to investigate the dependency of the estimated norm value on the background-subtracted region, we made smaller or larger the result of this fully automated procedure [32] by one to three voxels in all directions and relaunched the estimation procedure on the new area.

Robustness assessment

We firstevaluated the robustness of the SPEQTACLE algorithm. Robustness was defined as the ability of the automatic algorithm to provide consistent results for a given known object of

interest, considering varying image properties such as spatial sampling (voxel size), SNR, contrast, texture, filtering, etc. This evaluation was carried out using a dataset of phantoms containing homogeneous spheres on homogeneous background that were acquired in different PET/CT scanners, each with varyingacquisition and reconstruction parameters (see section D. Datasets).Homogeneous spheres on homogeneous backgroundare not appropriate for the evaluation of absolute accuracy since they represent a simplisticset-up and because of thebias due to cold sphere walls[34, 35]. On the other hand, they are well suited for the task of robustness estimationsince any present bias presentis the same for all acquisitions and they can provide a wide range in imaging settings for a given known object. The four spheres with largest diameters (37, 28, 22 and 17 mm) were segmented individually. The 13 and 10mm spheres were not included in the analysis because they were not filled in all acquisitions and are often too small with respect to the reconstructed voxel size to provide meaningfulresults.

Accuracy assessment

To evaluate the accuracy of the new algorithm relative to that of current state-of-the-art methods more challenging cases such as relatively large, complex-shaped and/or heterogeneous tumors were used considering both simulated realistic tumors and clinical tumor cases (see section D. datasets).

Evaluation metrics

For the robustness assessment, since the objects used are simple homogeneous spheres and the goal is to assess the consistency of results over various acquisitions of the same object and not absolute accuracy, the standard deviation of the determined volumes for a given sphere across the entire dataset (all scanners, all configurations) was reported as a measure of robustness.

For the accuracy evaluation, the classifications errors (CE) were used. In the simulated dataset, CE were calculated relatively to the known ground truth. In the clinical datasets, CE

were calculated relatively toa surrogate of truth obtained through a statistical consensus using the STAPLE (Simultaneous Truth And Performance Level Estimation) algorithm [36] applied to three manual delineations performed by experts with similar training and experience. CE may result from two contributions: the false negatives, the number of misclassified voxels within the ground truth, and the false positives, the number of misclassified voxels outside of the ground truth. CE as a percentage is then calculated as the sum of positive and negative misclassified voxels, divided by the number of voxels defining the ground truth[15]. CE were reported as mean±SD as well as with box-and-whisker plots in the figures.

Comparison with other methods

Within this evaluation framework, the proposed algorithm SPEQTACLE was compared to a couple of state-of-the-art methods which are improvement of the classical FCM: the Fuzzy Locally Adaptive Bayesian (FLAB) [16]and the Fuzzy Local Information C-means (FLICM)[30].Because the standard FCM has already been extensively evaluated and compared to these extensions or other previous segmentation approaches,including on PET images[15, 16, 37], it was not included in the present analysis.

FLAB combines a fuzzy measure with a Gaussian mixture model, and a stochastic estimation of the parameters from a FCM-based initialization. This method was developed initially for PET and thoroughly validated on both simulated and clinical datasets[16, 17, 23].FLICM is a recent FCM algorithm with a weighted norm taking into account outliers due to the noise[30]. This method uses two parameters: a regularization parameter and the size of the surrounding kernel. In the present work, we have set the parameter regularization equal to 1 and the kernel radius equal to 3 voxels, which are the recommended values[30] although they have not been optimized specifically for PET.

For all methods, the object of interest is first isolated in a 3D region of interest (ROI) containing the tumor, similarly as previously detailed for FLAB[15]. The number of

classes/clustersused was 2 for the robustness evaluation (homogeneous spheres) and 3 for the accuracy evaluation, in order to take into account potential tumor uptake heterogeneity. The two tumor classes were then unified for the error calculation with respect to the binary ground-truth (tumor/background).Thus, all algorithms were applied considering the same number of classes/clustersfor a given image.

The Wilcoxon rank sum testwas used to compare theresults between methods. P-values below 0.05 were considered significant.

D. Datasets

Homogeneous spheres phantoms

The dataset used for the robustness evaluation consists of NEMA phantoms containing spheres of various sizes (37, 28, 22, 17, 13, 10 mm)and filled with ^{18}F-FDG, thatwereacquired in three different PET/CT scanners: two PHILIPS scanners (a standard GEMINI and a time-of-flight (TOF) GEMINI), anda SIEMENSBiograph 16 scanner[8]. The standard iterative reconstruction algorithms associated with each scanner were used with their usual parameters: Time-of-Flight Maximum Likelihood-Expectation Maximization (TF ML-EM) for the GEMINI TOF, 3D Row Action Maximum Likelihood Algorithm (RAMLA) (2 iterations, relaxation parameter 0.05, Gaussian post-filteringwith 5mm FWHM) for the GEMINI, and Fourier rebinning (FORE) followed by Ordered Subsets Expectation Maximization (OSEM) (4 iterations, 8 subsets, Gaussian post-filtering with 5mm FWHM) for the Biograph16. All PET images were reconstructed using CT-based attenuation correction, as well as scatter and random coincidences. For each scanner, two different values for the following acquisition parameters and reconstruction settingswere considered: the contrast between the sphere and the background (4:1 and 8:1), the voxel size in the reconstruction matrix (2×2×2 and 4×4×4 or 5.33×5.33×2 mm^3) and the noise level (2 and 5 min of listmode data). Note that for the GEMINI acquisitions, the 28mm sphere was missingin the physical phantom. Figure 1 illustrates the images obtained for some of the acquisitions.

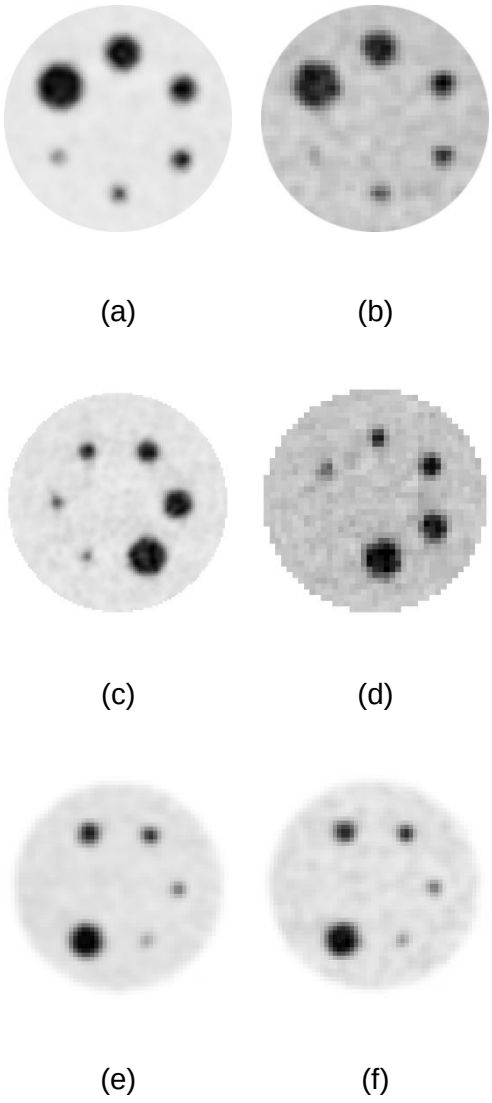


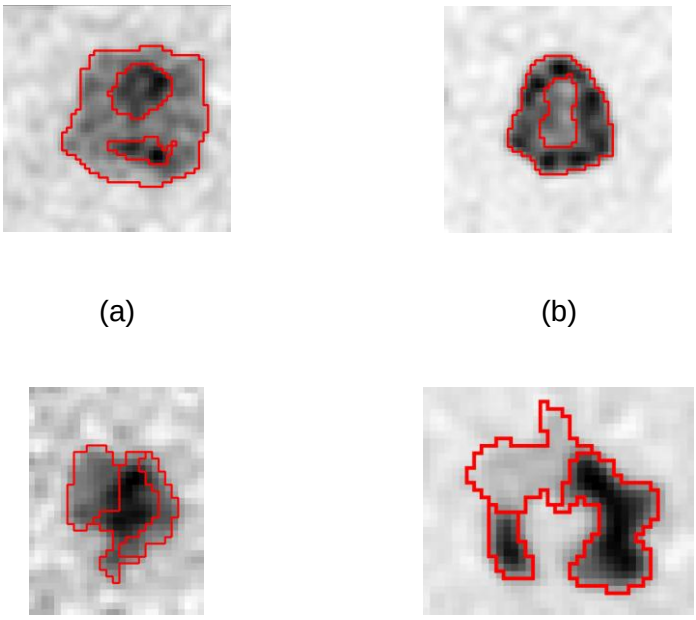
Fig 1. Examples of phantoms acquisitions: (a-b) the PHILIPS GEMINI TOF scanner with 5min acquisitionsand (a) ratio 8:1, voxels 2×2×2 mm^3, (b) ratio 4:1, 4×4×4 mm^3. (c-d) the SIEMENS scanner with 5min acquisitions and (c) ratio 8:1,voxels 2×2×2 mm^3, (d) ratio 4:1, 5.33×5.33×2 mm^3. (e-f) the PHILIPS GEMINI scanner with ratio 8:1, voxels 4×4×4 mm^3, and (e) 5min acquisition, (f) 2 min acquisition.

Simulated PET images

A set of 34 simulated PET tumor images with a wide range of contrast, noise levels, uptake heterogeneity and shape complexity was generated following a previously described methodology to obtain realistic complex shapes and uptake distributions of tumors for which

the exact ground-truth on a voxel-by-voxel basis is known[38, 39]. This dataset was built with relatively more challenging cases compared to previously conducted evaluations[16], in order to provide more complex tumor cases with combination of low SNR, high levels of heterogeneities and complex shapes. The important steps of the procedure used to generate these images is outlined below, and the reader is referred to[38, 39] for more details.

Each clinical tumor was first manually delineated on a clinical PET image by a nuclear medicine expert, thus creating a voxelized volume that represents the ground-truth of the tumor model used in the simulation. The activity levels attributed to each of the tumor parts were derived from the activity measured in the same areas of the tumor in the corresponding patient images. This ground-truth tumor structure was subsequently transformed into a Non-Uniform Rational B-Splines (NURBS) volume via Rhinoceros^{TM} (CADLINK software), for insertion into the NCAT phantom [40]attenuation maps at the same approximate position as located in the patient. No respiratory or cardiac motions were considered. Simulations using a model of the Philips PET/CT scanner previously validated with GATE (Geant4 Application for Tomography Emission) [41]were carried out. A total of 45 million coincidences were simulated corresponding to the statistics of a clinical acquisition over a single axial 18 cm field of view. Images were subsequently reconstructed using the One-Pass List mode Expectation Maximization (OPL-EM) (7 iterations, 1 subset). In some cases, the same 3D tumor shape was produced with different levels of contrast and heterogeneity, voxel sizes (4×4×4 and2×2×2 mm^3) and/or a different number of coincidences (45M or 20M) for different SNR realizations.Figure 2 illustrates some of the simulated tumors. The first two cases (fig. 2a-b) present relatively simpler shapes, higher contrast and SNR, whereas fig. 2c and 2d present more complex shapes and higher levels of noise and uptake heterogeneity.



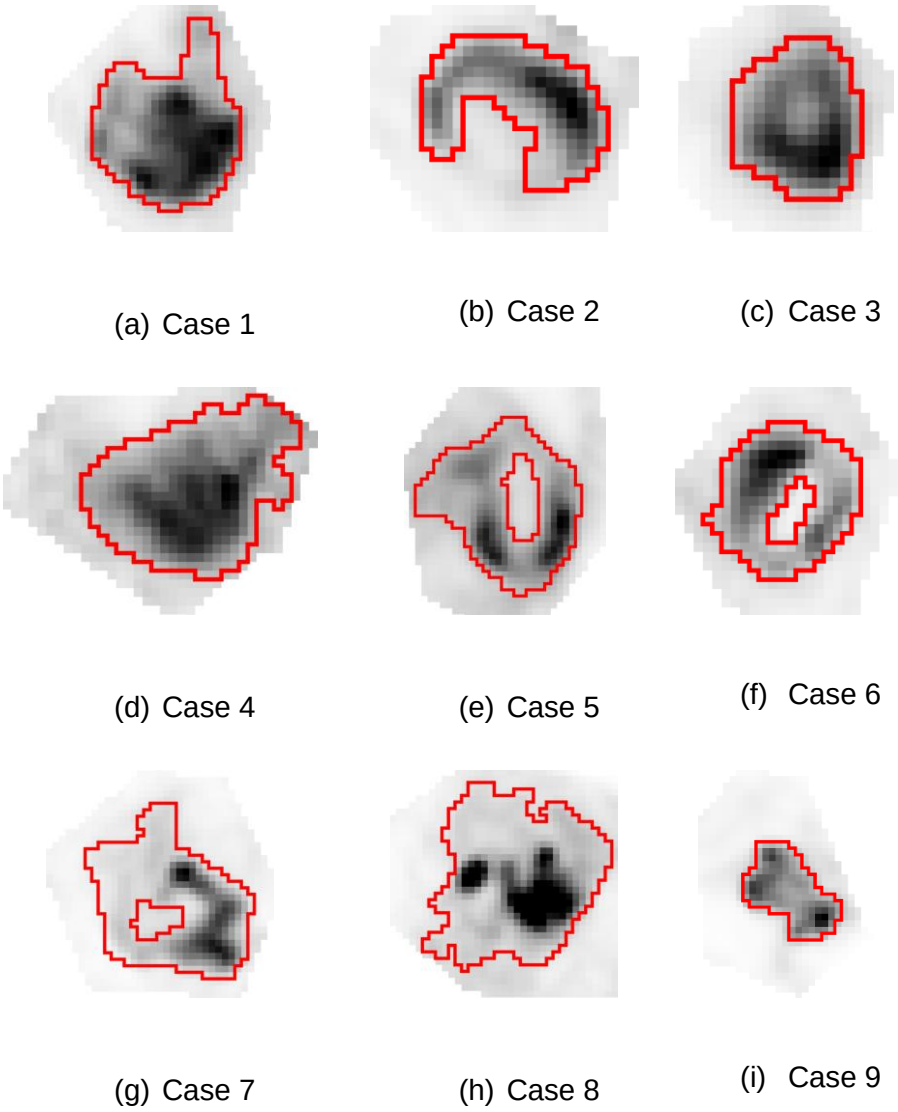
(a) (b)

(c) (d)

Fig 2. Four examples of simulated tumors. Red contours correspond to the simulation ground truth showing both external contours and sub-volumes heterogeneity.

Clinical PET images

Nine non-Small Cell Lung Cancer (NSCLC) tumors were chosen for their challenging nature with complex shapes and uptake heterogeneity.Patients fasted for at least 6 hours before 3D PET data was acquired on a Philips GEMINI PET/CT scanner without motion correction, 60±4 min after injection of 5MBq/kg of ^{18}F-FDG. Images were reconstructed with the 3D RAMLA algorithm (2 iterations, relaxation parameter 0.05, post-filtering with a Gaussian of 5 mm FWHM) and a voxel size of 4×4×4 mm^3, using CT-based attenuation correction, scatter and random correction[42]. In the absence of ground-truth for these volumes, 3 different experts delineated each tumor slice-by-slice with free display settings. A statistical consensus of the segmentations was then derived using the STAPLE algorithm to generate one surrogate of truth (fig. 3).



399 Fig 3. (a-i) Clinical images of 9 NSCLC tumors. Red contours correspond to the statistical
400 consensus of 3 different manual delineations.

401 ## Results

402 *Repeatability and dependency on initially selected tumor region*

403 The procedure was found perfectly repeatable with no variations in the resulting
404 segmentations on repeated applications to the same (previously defined) region of interest.
405 In addition, enlarging or reducing the size of the initial background-subtracted area by 1 to 3
406 voxels in all directions (equivalent to shrinking or increasing of the size of the region used to
407 estimate the norm by 5 to 15%) resulted in only minor variations in the estimated norm value

(3±11%, range –10% to +16%), and even smaller variations in the resulting segmentation (2±5%, range –4% to +7%). A substantial degradation of the segmentation results (20% difference) was observed when the reduction(area not covering sufficiently the tumor) or enlargement (too much background incorporated) of the initially estimated area exceeded 50%.

Robustness

The robustness of FLAB and standard FCMhas already been reported extensively[8]. In the current work we focused on three scanners and the 4 largest spheres, comparing SPEQTACLE to FLAB and FLICM.Figure4presents the robustness of each method, quantified bythe distributions of resulting volumes for each sphere as box-and-whisker plotsacross the entire dataset (3 scanners, all acquisition and reconstruction parameters). Although the accuracy was not under evaluation here, the true volume was also plotted for reference.

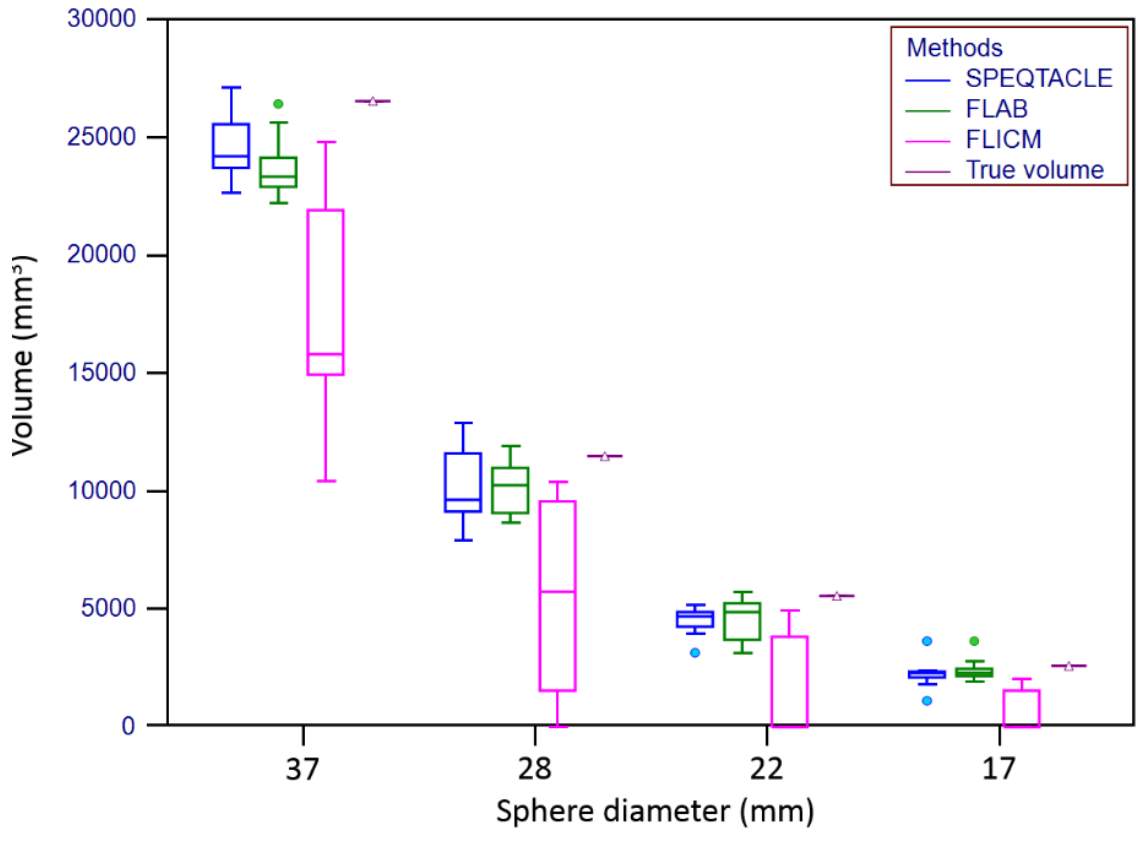
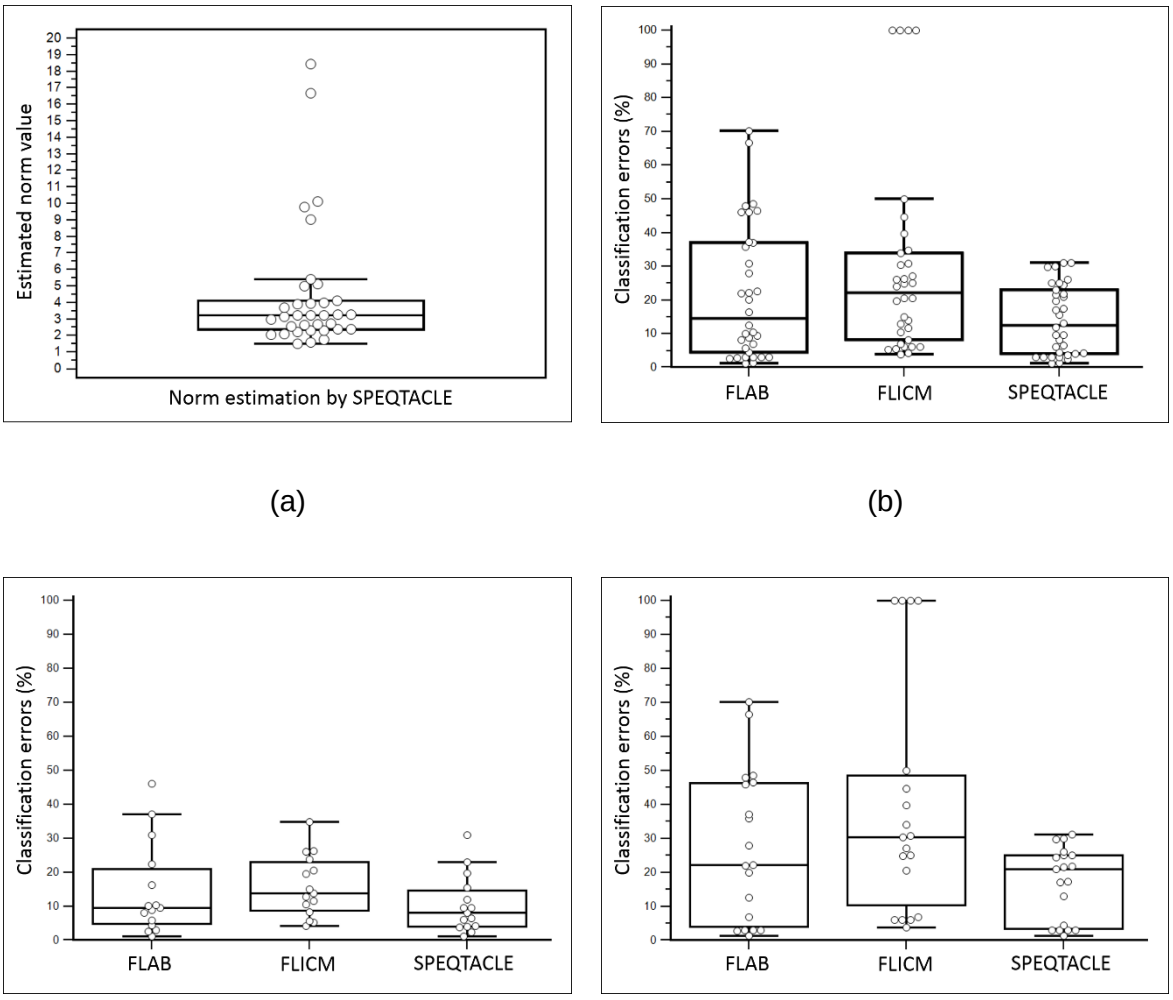


Fig 4. Distributions of volumes determined by the three methods under comparison for the four spheres of 37, 28, 22 and 17 mm in diameter across the entire robustness dataset. Box-and-whisker plots provide lower to upper quartile (25 to 75 percentile, central box), the median (middle line of the box) and the minimum to the maximum value, excluding "outlier values" which are displayed as separate dots.

The robustness performance of SPEQTACLE was satisfactory given the very large range of image characteristics. It was very similar and not statistically different ($p=0.15$) from FLAB with standard deviations of 5.4%, 16.9%, 12.7% and 26.6% for SPEQTACLE *vs.* 5.4%, 11.5%, 20.3%, and 19.3% for FLAB (for the 37, 28, 22 and 17mm spheres respectively). It should be emphasized that there were 2 outliers for the 17mm sphere and 1 for the 22 mm sphere (fig. 4). These were associated with images of some of the acquisitions for which the spheres were barely visible and spatially sampled with large voxels (see fig. 1b for an example), which explains the substantial deviation observed for these specific cases. When excluding these outliers, the robustness of SPEQTACLE increased with lower standard deviations of 7.9% and 18.8% for the 22 and 17mm sphere respectively.

FLICM exhibited significantly lower robustness ($p<0.0001$) than FLAB and SPEQTACLE. For the spheres 28, 22 and 17 mm, this was mostly due to segmentation failures in several cases for sphere diameters ≤28 mm, with the segmentation filling the entire ROI leading to extremely large volumes. For these complete failures, we limited the resulting volume to twice the expected volume of the sphere, leading to standard deviations of 68.9%, 40.9% and 43.7% for the spheres of 28, 22 and 17mm respectively. However for the largest sphere (37 mm in diameter), the standard deviation was also higher (26.8%) than SPEQTACLE and FLAB, without an associated segmentation failure, but rather very different results depending on the different image characteristics considered.

447 ***Accuracy***



448 Fig. 5 (a) Box-and-whisker plot of the norm parameter estimated by SPEQTACLE for the
449 entire set of simulated PET images. (b-d) Comparison of error rates for the three methods
450 with box-and-whisker plots, for (b) the 34 simulated tumors PET images, (c) the subset of
451 cases with estimated norm<3 and (d) cases with norm>3.

452 Figure5a shows the distribution of the values for the norm parameter as estimated by
453 SPEQTACLE. We recall that a value of 2corresponds to the standard FCM case. Almost half
454 the cases considered had an estimated norm between 3 and 6. Five cases led to estimated
455 norm values of 9 to 19. Given this distribution, we report the accuracy for the entire dataset,
456 then for the subset of cases with norm<3 (15 cases) and finally for >3 (19 cases), as we can

457 reasonably expect a larger improvement using SPEQTACLE over the two other algorithms
458 for higher norm values.

459 Figure 5bshows the classification errorsresults obtained by the threemethods under
460 comparison, for the entire set of 34 images.SPEQTACLE was found to provide lower CE
461 than FLAB ($p=0.0044$) and FLICM ($p<0.0001$). FLAB, FLICM and SPEQTACLE led to CE of
462 21.8±19.8% (median 14.5%, range 1.2 – 70.2%), 29±29% (median 22.3%, range 3.9 –
463 100.0%) and 14.4±10.6% (median 12.5%, range 1.3 – 37.9%)respectively. No errorsabove
464 40% were observed for SPEQTACLE contrary to FLAB (up to 50-70% errors) and FLICM
465 that even had four cases with >100% errors (complete failure of the segmentation, CE limited
466 to 100%). SPEQTACLE had more cases with errors below 10% and between 10% and 20%
467 than FLAB and FLICM, and fewer cases with errors between 20% and 50%.

468 Figure 5c provides the classification errors for the 15 images for which the estimated norm
469 was <3.In this first subset, although SPEQTACLE led to the best results (10.5±8.5%, median
470 8.3%, range 1.3 – 31%) with significantly lower errors than FLICM (15.3±9.1%, median
471 12.9%, range 4.2 – 34.8%, $p=0.0215$), no significant differences were found between
472 SPEQTACLE and FLAB (14.5±13.6%, median 9.5%, range 1.2 – 46.1%, $p=0.22$). No errors
473 above 50% were observed for any method.It should be emphasized that despite differences
474 between the three methods, all three achieved high accuracy performance with <20% CE for
475 the majority of cases.

476 Figure 5dprovides the classification errors for the second subset of 19 images for which the
477 estimated norm was >3.In this dataset of clearly more challenging cases, with an error rate of
478 17.4±11.3% (median 21%, range 1.4 – 37.9%), SPEQTACLE significantly outperformed all
479 other methods:FLAB with 27.6±22.2% (median 22.2%, range 1.4 – 70.2%) ($p=0.0092$) and
480 FLICM with 39.9±34.6% (median 30.5%, range 3.9 – 100.0%) ($p<0.0001$).No errors above
481 50% were observed for SPEQTACLE contrary to FLAB and FLICM, and there were less
482 errors between 20 and 50% for SPEQTACLE than for FLAB and FLICM. Overall, the

accuracy achieved by SPEQTACLE in this dataset of very challenging cases was satisfactory, with a maximum CE below 38% and a mean of17%. Figure 6 provides some visual examples of segmentation results for the simulated tumors.

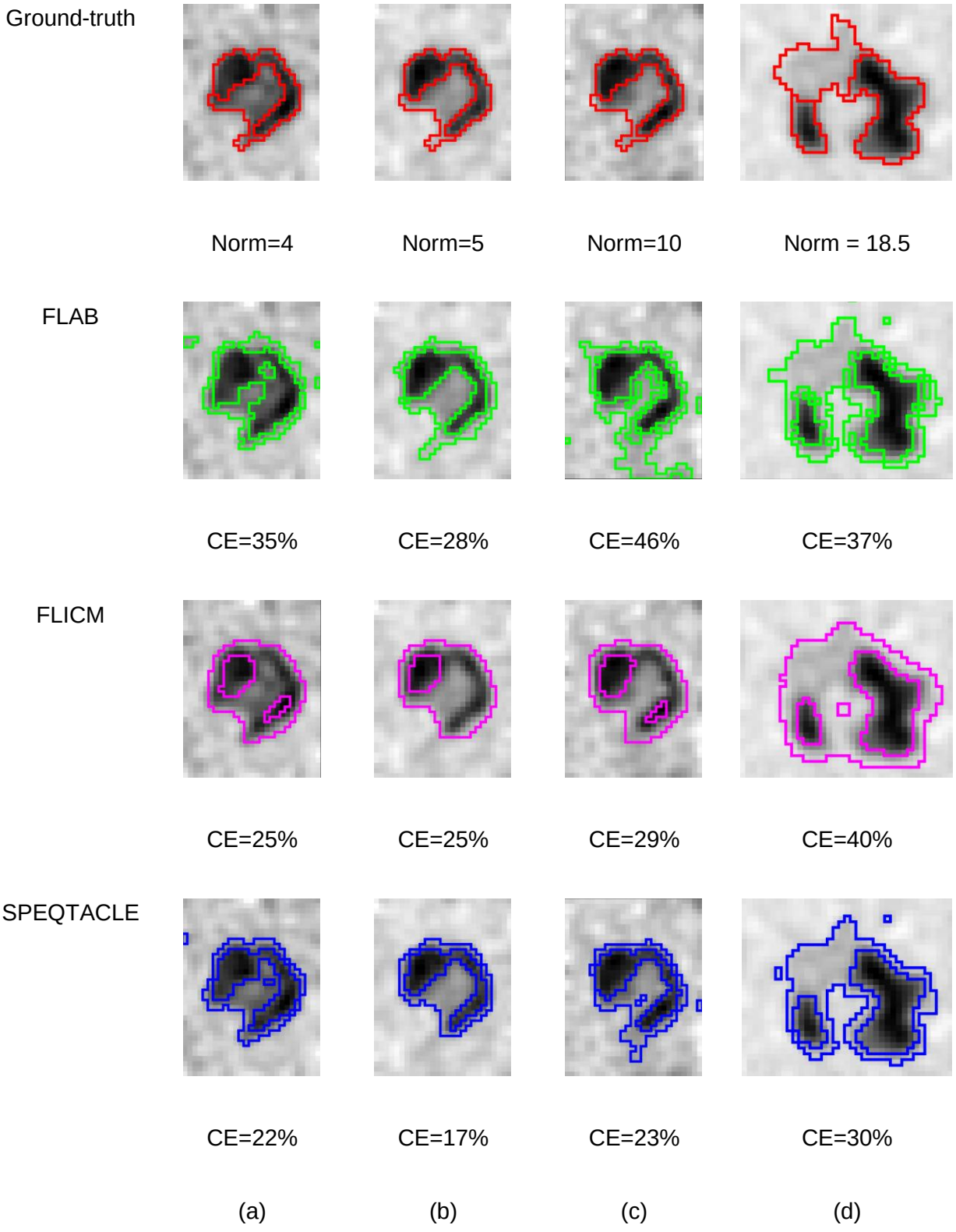
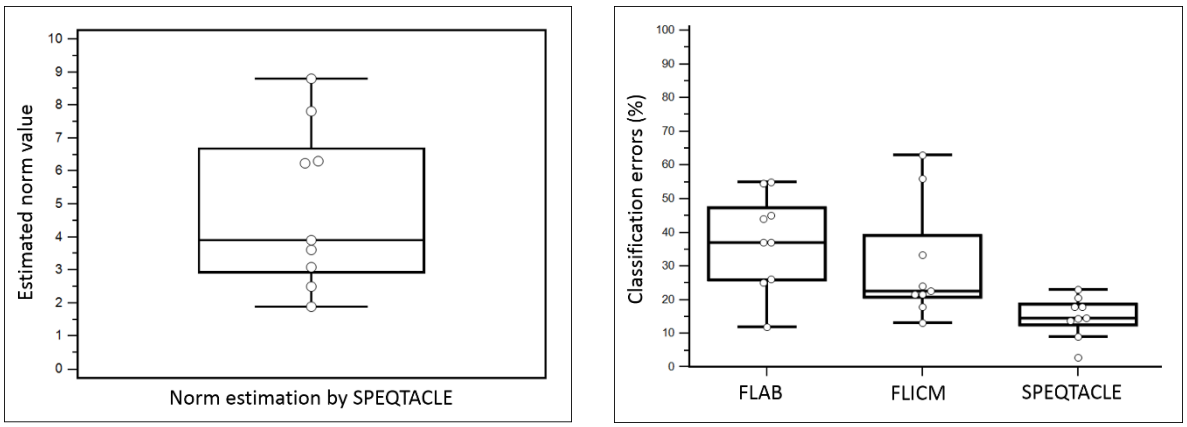


Fig 6. Segmentation results for (a-c) the same simulated tumor with increasing complexity: combinations of noise levels and heterogeneity both within the tumor (contrast between the various sub-volumes of the tumor) or in terms of overall contrast between the tumor and the background. These configurations were found to correspond to increasing estimated norm values: (a) 4, (b) 5 and (c) 10. (d)presents a tumor with complex shape and high levels of heterogeneity for which the norm was estimated at 18.45. First row is ground-truth (red) whereas second, third and fourth rows are results from FLAB (green), FLICM (magenta) and SPEQTACLE (blue).

Figure 7shows the estimated norm values (fig. 7a) and the classification errors(fig. 7b) for the nine clinical images.Norm values estimated by SPEQTACLE were between 2 and 9, with most of them being >3 (7 out of 9 cases). The best performance was obtained with SPEQTACLE with significantly ($p<0.004$) lower errors (mean 14.9±6.1%, range 2.9 – 23%) with respect to the STAPLE-derived consensus of manual delineations, compared to FLAB (mean 37.3±14.3%, range 12 – 55%) and FLICM (30.4±17.4%, range 13.2 – 63%).



(a) (b)

Fig 7. (a) Box-and-whisker plot of the norm parameter estimated by SPEQTACLE and (b) CE for the three methods, for the clinical dataset.

Figure 8 shows the results of segmentation for all 9 clinical cases.For cases 3 and 9, the three methods led to similar results, as the level of heterogeneity is relatively lower with

respect to the high overall contrast between the tumor and the surrounding background. On the one hand, for cases 1, 4, 5 and 6, it was observed that FLAB underestimated the spatial extent selected by the experts, by focusing on the high intensity uptake region, whereas FLICM led to results closer to the manual contours. On the other hand, for cases 2, 7 and 8, on the contrary FLAB slightly overestimated the manual contours, whereas FLICM underestimated it, missing the large areas with lower uptake. In all cases, SPEQTACLE demonstrated higheraccuracywith results closer to the manual delineations.

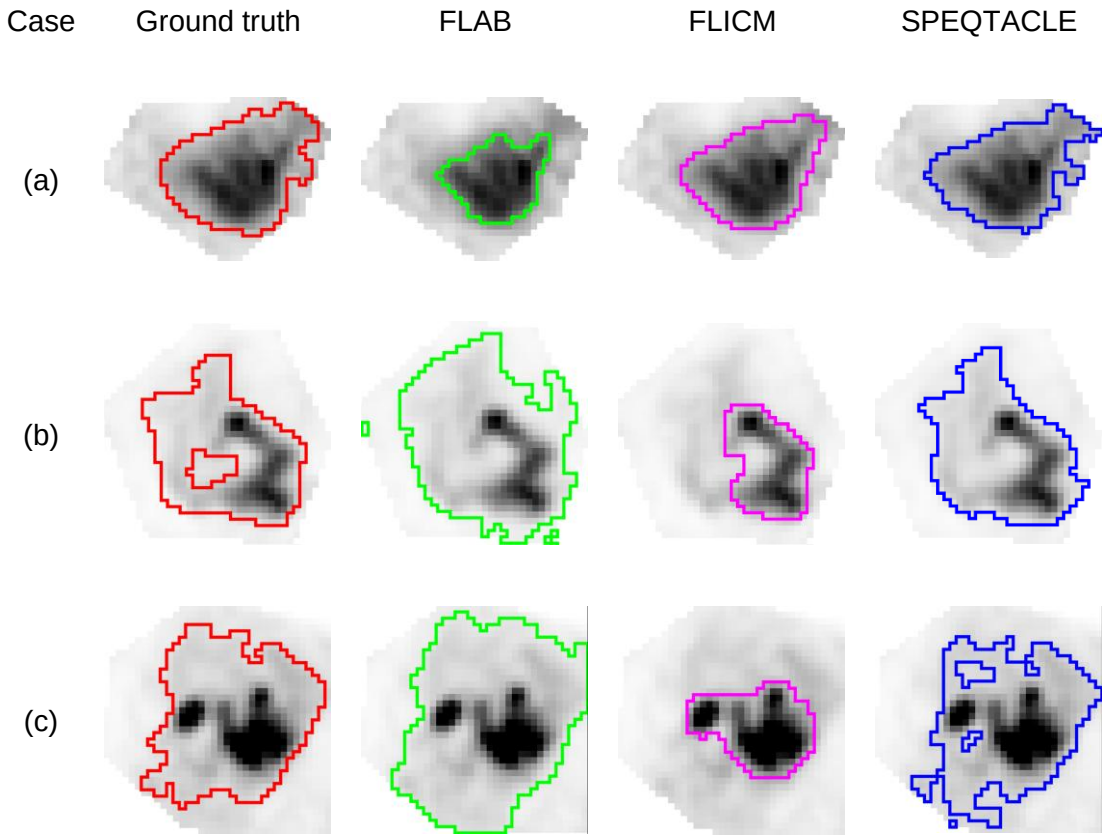


Fig 8. Examples of delineationsfor clinical cases (a) 4, (b) 7 and (c) 8 from Fig. 3 (d), (g) and (h): consensus of manual (red), FLAB (green), FLICM (magenta) and SPEQTACLE (blue).

Discussion

Although promising results for PET tumor delineation in a realistic setting beyond the validation using simple cases (spherical and/or homogeneous uptakes)have been recently achieved by several methods[11]there is still room for improvement, particularly in the case of

highly heterogeneous and complexshapes.The use of the fuzzy C-means clustering algorithm for delineation of PET tumors has been considered previouslyshowing a limited performance both in accuracy [15, 43, 44]and robustness[8]. Among the recent methods dedicated to PET that demonstrated promising accuracy, the fuzzy C-means algorithm was improved using a rather complex pipelinecombining spatial correlation modeling and pre-processing in the wavelet domain [44]. In the presented work, we rather focused on the generalization and full automation of the FCM approach to improve its accuracy and its ability to deal with challenging and complex PET tumor images, by implementing an estimation of the norm on a case-by-case basis. The improved accuracy results that we obtained on the validation datasets suggest that the optimal norm parameter can indeed be different for each PET tumor image and can vary substantially across cases, making anautomatic estimation essential in the accuracy of the FCM segmentation results.

It should be emphasized that SPEQTACLE did not undergo any pre-processing or pre-optimization and that no parameter was set or chosen to optimize the obtained results on the evaluation datasets (either phantoms, realistic simulated or clinical tumors).The improved accuracy that SPEQTACLE achieved is thereforeentirely due to its automatic estimation framework and its associated ability to adapt its norm parameter to varying properties of the image. The advantage of SPEQTACLE compared to other fuzzy clustering-based methods such as FLAB or FLICM thus lies on its ability to estimate reliably the norm parameter value on a case-by-case basis. In addition, the proposed norm estimation scheme is deterministic and convergent, therefore the repeatability of the algorithm was found to be perfect with zero variability in the results on repeated segmentations of the same image, which is an important point to ensure clinical acceptance for use by the physicians. In addition, the estimation of the norm was also found to be robust with respect to slightly larger or smaller initial determination of the tumor class using a background-subtraction approach[32]. In order to reach substantial differences in the segmentation results, this area had to be enlarged or

shrunk by more than 50%, which is very unlikely to occurunless highly inaccurate methods are used to define the initial region.

We showed that SPEQTACLE led to significantlyhigher accuracyin delineating tumor volumes with higher complexity (either in terms of shape, heterogeneity, noise levels and/or contrast), associated with a norm value higher than 3, on both simulated and clinical datasets. On the other hand, for simpler objects of interest (norm value below 3), we found that SPEQTACLE provided similar (although slightly improved) accuracy as FLAB and FLICM.Given the improved accuracy obtained with respect to FLAB on a dataset with a large range of contrast and noise levels as well as heterogeneity and shape, we expected that the robustness of SPEQTACLE should be at least similar as the one of FLAB. We indeed confirmed through a robustness analysis that the proposed automatic norm estimation scheme does not lead to decreased robustness with respect to varying image properties associated with the use of different PET/CT scanner models, reconstruction algorithms, or acquisition and reconstruction settings. Indeed, the level of robustness exhibited by SPEQTACLE was found to be similar to the one of FLAB, which had already been demonstrated as substantially more robust than standard FCM[8]. FLICM however was found to be much less robust, with segmentation failures for some of the configurations in the dataset. Given the fact that FLICM performed reasonably well on the accuracy dataset, its failure on the robustness evaluationmight be due to the two parameters (the regularization parameter and the size of the surrounding kernel) that were set *a priori* in this study using recommended values that might not be appropriate for some of the PET images of the robustness dataset. The overall performance of FLICM might therefore be improved by optimizing these two parameters for each phantom acquisition, which is however out of the scope of the present work.

From a clinical point of view, our method might be easier than most of the previously proposed onesto implement in a clinical setting because it is fully automatic and perfectly repeatable, with no user intervention for parameterization beyond the localization of the

tumor in the whole-body image and its isolation in a 3D ROI. It is also very fast due to its low computational cost; thesegmentation of thelargest tumor (55×55×25 voxels) requires less than 1 min on a standard computer (CPU E5520 2.27 GHz×8), which could be easily shortened through algorithmic optimization and parallel computing or GPU implementation. Moreover, the algorithm itselfuses a negligible amount of memory.

The present work has a few limitations. It should be reminded that the proposed algorithm aims at the accurate delineation of a single pathological uptake previously detected and isolated in a ROI, similarly as FLAB. It was therefore not evaluated within the context of the simultaneous segmentation of multiple tumors (as each tumor should be processed independently when using SPEQTACLE), the detection of tumors and/or lymph nodes in a whole-body image[45], nor the segmentation of diffuse and multifocal uptakes such as in pulmonary infection[46]. Also, we did not investigate the impact on the resulting segmentation of theinitial ROI selection, which is a first step as in most of published methods for PET tumor delineation[10, 11]. However, we already showed that this step has a very limited impact on the results for FLAB, as long as the ROI selection is made without incorporating nearby non-relevant uptake that would bias the estimation process[15]. Given that SPEQTACLE demonstrated similar robustness as FLAB, the impact of this step should be similarly low. Second, we did not include a large number of methods to compare SPEQTACLE with. Given its previous validation and demonstrated performance, FLAB can be considered a state-of-the-art method and our primary goal was to improve on that approach for challenging cases. A full comparison with numerous other methods was out of the scope of this work and might be conducted in the future using the benchmark currently being developed by the AAPM taskgroup 211[147]. Second, the robustness analysis was carried out on a smaller dataset than for the previously reported analysis for FLAB, FCM and thresholding methods[8], however the dataset is certainly representative enough to provide a clear picture. Third, we did not evaluate the algorithms on clinical datasets with histopathology associated measurements.

[1] http://aapm.org/org/structure/default.asp?committee_code=TG211

The one dataset available to us consists of maximum diameter measurements only[17], which might not be sufficient to highlight differences between the advanced algorithms under comparison. On the other hand, a benchmark developed by the AAPM Taskgroup 211 is expected to contain several clinical datasets with histopathological volumes[47], and could be used for future comparison studies. Finally, in the present implementation, the norm parameter was estimated from an automatically pre-segmented estimation of the tumor region, using a background-subtraction approach [32] in order to obtain a first guess of the tumor class. The estimated norm was then used for all classes in the segmentation.In future work, it would therefore be possible topotentially improve the algorithm performanceby estimating a norm parameter for each class in the ROI. In this case, the minimized criterion in GFCM becomes:

$$\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\left|y_u-\mu_i\right|^{\alpha_i} .$$

The norm parameter α_i cannot be estimated by using the Newton-Raphson algorithm on the minimized criterion of equation (4). Indeed, the norm parameter is essentially dependent on the statistical behavior of the data and generally there is no solution $\alpha_i \neq 1$ which minimizes equation (4). Thus minimizing equation (4) according to α_i is equivalent to solving:

$$\sum_{u\in V}\sum_{i=1}^{C} p_{u,i}^{m}\log\left(\left|y_u-\mu_i\right|\right)\left|y_u-\mu_i\right|^{\alpha_i}=0.$$

Consequently, it depends on how data are scaled and the presence of y_u such that $\left|y_u-\mu_i\right|>1$ contributes to making this derivative > 0. Amongst other possible extensions, it will be interesting to estimate a variance parameter additionally to the center parameter μ_i and the norm parameter. Such a method would be able to fit more completely the statistical distribution of the intensities. Indeed, μ_i controls the mean of intensities for each cluster, α_i controls the shape of the distribution whereas the variance parameter controls the disparity

of each cluster.Another future work will consist in extending SPEQTACLE to the multimodal situation for which each voxel becomes a vector whose components represent intensities taken from each image modality, for instance PET, CT and MRI. In the multimodal version, a norm parameter has to be estimated for each modality. The minimized criterion will thus have the same form by replacing the absolute value by a sum of absolute values.

Conclusions

In this paper, we have presented a fully automatic method for estimating the norm parameter in a generalized fuzzy C-meansframework. We have developed and validated this new methodfor PET tumor delineation, andnamed it SPEQTACLE for *Spatial Positron Emission Quantification of Tumor:AutomatiCLp-norm Estimation*. The proposed approach is fully automated and perfectly repeatable. It provides improved accuracy with respect to state-of-the art methods for realistic challenging delineation cases. Thiswas demonstrated on both simulated and clinical datasets with complex shapes, high levels of uptake heterogeneity. The improvement in accuracy was achieved without sacrificing robustness *vs.* varying image properties in a multi-centric setting, which is crucial if the method is to be widely applicable in clinical practice. Future extensions of SPEQTACLE will include a multimodal version of the algorithm for PET/CT, PET/MRI and other multimodal medical imaging applications, as well as a multi-class norm estimation scheme to improve the algorithm performance.

References

[1] S. Hess, B.A. Blomberg, H.J. Zhu, P.F. Høilund-Carlsen, and A. Alavi, "The Pivotal Role of FDG-PET/CT in Modern Medicine," Acad. Radiol. **21**(2), 232–249 (2014).

[2] G.C. Pereira, M. Traughber, and R.F. Muzic, "The role of imaging in radiation therapy planning: past, present, and future," BioMed Res. Int. **2014**, 231090 (2014).

[3] K. Herrmann, M.R. Benz, B.J. Krause, K.L. Pomykala, A.K. Buck, and J. Czernin, "(18)F-FDG-PET/CT in evaluating response to therapy in solid tumors: where we are and where we can go," Q J Nucl Med Mol Imaging **55**(6), 620–32 (2011).

[4] T. Carlier and C. Bailly, "State-Of-The-Art and Recent Advances in Quantification for Therapeutic Follow-Up in Oncology Using PET," Front. Med. **2**, 18 (2015).

[5] M.K. Rahim, S.E. Kim, H. So, H.J. Kim, G.J. Cheon, E.S. Lee, K.W. Kang, and D.S. Lee, "Recent Trends in PET Image Interpretations Using Volumetric and Texture-based Quantification Methods in Nuclear Oncology," Nucl. Med. Mol. Imaging **48**(1), 1–15 (2014).

[6] J.P.B. O'Connor, C.J. Rose, J.C. Waterton, R.A.D. Carano, G.J.M. Parker, and A. Jackson, "Imaging Intratumor Heterogeneity: Role in Therapy Response, Resistance, and Clinical Outcome," Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. **21**(2), 249–257 (2015).

[7] S. Houshmand, A. Salavati, S. Hess, T.J. Werner, A. Alavi, and H. Zaidi, "An update on novel quantitative techniques in the context of evolving whole-body PET imaging," PET Clin. **10**(1), 45–58 (2015).

[8] M. Hatt, C. Cheze Le Rest, N. Albarghach, O. Pradier, and D. Visvikis, "PET functional volume delineation: a robustness and repeatability study," Eur J Nucl Med Mol Imaging **38**(4), 663–72 (2011).

[9] R. Boellaard, M.J. O'Doherty, W.A. Weber, F.M. Mottaghy, M.N. Lonsdale, S.G. Stroobants, W.J. Oyen, J. Kotzerke, O.S. Hoekstra, J. Pruim, P.K. Marsden, K. Tatsch, C.J. Hoekstra, E.P. Visser, B. Arends, F.J. Verzijlbergen, J.M. Zijlstra, E.F. Comans, A.A. Lammertsma, A.M. Paans, A.T. Willemsen, T. Beyer, A. Bockisch, C. Schaefer-Prokop, D. Delbeke, R.P. Baum, A. Chiti, and B.J. Krause, "FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0," Eur J Nucl Med Mol Imaging **37**(1), 181–200 (2010).

[10] M. Hatt, N. Boussion, C. Cheze-Le Rest, D. Visvikis, and O. Pradier, "[Metabolically active volumes automatic delineation methodologies in PET imaging: review and perspectives]," Cancer Radiother **16**(1), 70–81; quiz 82, 84 (2012).

[11] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D.J. Mollura, "A review on segmentation of positron emission tomography images," Comput. Biol. Med. **50**, 76–96 (2014).

[12] B. Braathen, W. Pieczynnski, and P. Masson, *Global and local methods of unsupervised Bayesian segmentations of images*, Mach. Graph. Vis. 39–52 (1993).

[13] D. Benboudjema and W. Pieczynski, "Unsupervised statistical segmentation of nonstationary images using triplet Markov fields," IEEE Trans Pattern Anal Mach Intell **29**(8), 1367–78 (2007).

[14] M. Hatt, F. Lamare, N. Boussion, A. Turzo, C. Collet, F. Salzenstein, C. Roux, P. Jarritt, K. Carson, C. Cheze-Le Rest, and D. Visvikis, "Fuzzy hidden Markov chains segmentation for volume determination and quantitation in PET," Phys Med Biol **52**(12), 3467–91 (2007).

[15] M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, and D. Visvikis, "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET," IEEE Trans Med Imaging **28**(6), 881–93 (2009).

[16] M. Hatt, C. Cheze le Rest, P. Descourt, A. Dekker, D. De Ruysscher, M. Oellers, P. Lambin, O. Pradier, and D. Visvikis, "Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications," Int J Radiat Oncol Biol Phys **77**(1), 301–8 (2010).

[17] M. Hatt, C. Cheze-le Rest, A. van Baardwijk, P. Lambin, O. Pradier, and D. Visvikis, “Impact of tumor size and tracer uptake heterogeneity in (18)F-FDG PET and CT non-small cell lung cancer tumor delineation,” J Nucl Med **52**(11), 1690–7 (2011).

[18] M. Hatt, C. Cheze-Le Rest, E.O. Aboagye, L.M. Kenny, L. Rosso, F.E. Turkheimer, N.M. Albarghach, J.P. Metges, O. Pradier, and D. Visvikis, “Reproducibility of 18F-FDG and 3’-deoxy-3’-18F-fluorothymidine PET tumor volume measurements,” J Nucl Med **51**(9), 1368–76 (2010).

[19] B.H. de Figueiredo, M. Antoine, R. Trouette, P. Lagarde, A. Petit, F. Lamare, M. Hatt, and P. Fernandez, “Use of FDG-PET to guide dose prescription heterogeneity in stereotactic body radiation therapy for lung cancers with volumetric modulated arc therapy: a feasibility study,” Radiat. Oncol. Lond. Engl. **9**, 300 (2014).

[20] B. Henriques de Figueiredo, C. Zacharatou, S. Galland-Girodet, J. Benech, H. De Clermont-Gallerande, F. Lamare, M. Hatt, L. Digue, E. De Mones Del Pujol, and P. Fernandez, “Hypoxia imaging with [18F]-FMISO-PET for guided dose escalation with intensity-modulated radiotherapy in head-and-neck cancers,” Strahlenther. Onkol. Organ Dtsch. Rontgengesellschaft Al (2014).

[21] A.I.J. Arens, E.G.C. Troost, B.A.W. Hoeben, W. Grootjans, J.A. Lee, V. Grégoire, M. Hatt, D. Visvikis, J. Bussink, W.J.G. Oyen, J.H.A.M. Kaanders, and E.P. Visser, “Semiautomatic methods for segmentation of the proliferative tumour volume on sequential FLT PET/CT images in head and neck carcinomas and their relation to clinical outcome,” Eur. J. Nucl. Med. Mol. Imaging **41**(5), 915–924 (2014).

[22] M. Hatt, A.L. Maitre, D. Wallach, H. Fayad, and D. Visvikis, “Comparison of different methods of incorporating respiratory motion for lung cancer tumor volume delineation on PET images: a simulation study,” Phys Med Biol **57**(22), 7409–30 (2012).

[23] A. Le Maitre, M. Hatt, O. Pradier, C. Cheze-le Rest, and D. Visvikis, “Impact of the accuracy of automatic tumour functional volume delineation on radiotherapy treatment planning,” Phys Med Biol **57**(17), 5381–97 (2012).

[24] D.E. Gustafson and W.C. Kessel, “Fuzzy clustering with a fuzzy covariance matrix,” in *1978 IEEE Conf. Decis. Control 17th Symp. Adapt. Process.*(1978), pp. 761–766.

[25] R.J. Hathaway, J.C. Bezdek, and Y. Hu, “Generalized Fuzzy C-means Clustering Strategies Using Lp Norm Distances,” Trans Fuz Sys **8**(5), 576–582 (2000).

[26] S. Chen and D. Zhang, “Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure,” IEEE Trans. Syst. Man Cybern. Part B Cybern. **34**(4), 1907–1916 (2004).

[27] M.-S. Yang and K.-L. Wu, “Unsupervised Possibilistic Clustering,” Pattern Recogn **39**(1), 5–21 (2006).

[28] M. Daniel, “Belief Functions: A Revision of Plausibility Conflict and Pignistic Conflict.,” in *SUM*, edited by W. Liu, V.S. Subrahmanian and J. Wijsen (Springer, 2013), pp. 190–203.

[29] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag, and T. Moriarty, “A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data,” IEEE Trans. Med. Imaging **21**(3), 193–199 (2002).

[30] S. Krinidis and V. Chatzis, “A robust fuzzy local information C-Means clustering algorithm,” IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **19**(5), 1328–1337 (2010).

[31] I. Gath and A. Geva, “Unsupervised optimal fuzzy clustering,” IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 773–780 (1989).

[32] I.A. Burger, H.A. Vargas, B.J. Beattie, D.A. Goldman, J. Zheng, S.M. Larson, J.L. Humm, and C.R. Schmidtlein, “How to assess background activity: introducing a histogram-based analysis as a first step for accurate one-step PET quantification,” Nucl. Med. Commun. **35**(3), 316–324 (2014).

[33] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry* (American Mathematical Society, 2007).

[34] B. Berthon, C. Marshall, A. Edwards, M. Evans, and E. Spezi, “Influence of cold walls on PET image quantification and volume segmentation: a phantom study,” Med. Phys. **40**(8), 082505 (2013).

[35] F. Hofheinz, S. Dittrich, C. Pötzsch, and J. van den Hoff, "Effects of cold sphere walls in PET phantom measurements on the volume reproducing threshold," Phys. Med. Biol. **55**(4), 1099–1113 (2010).

[36] S.K. Warfield, K.H. Zou, and W.M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," IEEE Trans Med Imaging **23**(7), 903–21 (2004).

[37] H. Zaidi, M. Abdoli, C.L. Fuentes, and I.M. El Naqa, "Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma," Eur. J. Nucl. Med. Mol. Imaging **39**(5), 881–891 (2012).

[38] A. Le Maitre, W. Segars, S. Marache, A. Reilhac, M. Hatt, S. Tomei, C. Lartizien, and D. Visvikis, "Incorporating Patient-Specific Variability in the Simulation of Realistic Whole-Body 18F-FDG Distributions for Oncology Applications," Proc. IEEE **9**(12), 2026–2038 (2009).

[39] P. Papadimitroulas, G. Loudos, A. Le Maitre, M. Hatt, F. Tixier, N. Efthimiou, G.C. Nikiforidis, D. Visvikis, and G.C. Kagadis, "Investigation of realistic PET simulations incorporating tumor patient's specificity using anthropomorphic models: creation of an oncology database," Med. Phys. **40**(11), 112506 (2013).

[40] W. Segars, Development and Application of the New Dynamic NURBS-based Cardiac-Torso (NCAT) phantom (2001).

[41] F. Lamare, A. Turzo, Y. Bizais, C.C. Le Rest, and D. Visvikis, "Validation of a Monte Carlo simulation of the Philips Allegro/GEMINI PET systems using GATE," Phys Med Biol **51**(4), 943–62 (2006).

[42] D. Visvikis, A. Turzo, A. Gouret, P. Damine, F. Lamare, Y. Bizais, and C. Cheze Le Rest, "Characterisation of SUV accuracy in FDG PET using 3-D RAMLA and the Philips Allegro PET scanner," J. Nucl. Med. **45**(5), 103 (2004).

[43] D.C. Weber, H. Wang, L. Cozzi, G. Dipasquale, H.G. Khan, O. Ratib, M. Rouzaud, H. Vees, H. Zaidi, and R. Miralbell, "RapidArc, intensity modulated photon and proton techniques for recurrent prostate cancer in previously irradiated patients: a treatment planning comparison study," Radiat Oncol **4**, 34 (2009).

[44] S. Belhassen and H. Zaidi, "A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET," Med Phys **37**(3), 1309–24 (2010).

[45] L. Bi, J. Kim, L. Wen, and D.D. Feng, "Automated and robust PERCIST-based thresholding framework for whole body PET-CT studies," Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Conf. **2012**, 5335–5338 (2012).

[46] B. Foster, U. Bagci, null Ziyue Xu, B. Dey, B. Luna, W. Bishai, S. Jain, and D.J. Mollura, "Segmentation of PET images for computer-aided functional quantification of tuberculosis in small animal models," IEEE Trans. Biomed. Eng. **61**(3), 711–724 (2014).

[47] T. Shepherd, B. Berthon, P. Galavis, E. Spezi, A. Apte, J. Lee, D. Visvikis, M. Hatt, E. de Bernardi, S. Das, I. El Naqa, U. Nestle, C. Schmidtlein, H. Zaidi, and A. Kirov, "Design of a benchmark platform for evaluating PET-based contouring accuracy in oncology applications," Eur. J. Nucl. Med. Mol. Imaging **39**, S264–S264 (2012).

Appendices

A. FCM minimization step

The minimization process for FCM is achieved recursively until convergence:

1. Let $p_{u,i}^{(0)}$ and $\mu_i^{(0)}$ be initial values;
2. From $p_{u,i}^{(q)}$ compute: $\mu_i^{(q+1)} = \dfrac{\sum_{u \in V} \left(p_{u,i}^{(q)}\right)^m y_u}{\sum_{u \in V} \left(p_{u,i}^{(q)}\right)^m}$
3. From $\mu_i^{(q+1)}$ compute: $p_{u,i}^{(q+1)} = \dfrac{1}{\sum_{k=1}^{C} \left(\dfrac{\left| y_u - \mu_i^{(q+1)} \right|}{\left| y_u - \mu_k^{(q+1)} \right|} \right)^{\frac{2}{m-1}}}$.

B. Newton-Raphson algorithm

Let f be a derivable function from $\Re$ to $\Re$, the Newton-Raphson algorithm is an algorithm to find the solution a such that $f(a) = 0$. The Newton-Raphson works as following:

1. Set a_0 an initial value;
2. $a_{n+1} = a_n - \dfrac{f(a_n)}{f'(a_n)}$.

C. GFCM minimization step (norm parameter is known)

For fixed norm parameter α and weight parameters $(p_{u,i})_{1 \le i \le C}$, the center μ_j is estimated by minimizing:

$\sum_{u \in V} p_{u,j}^m \left| y_u - \mu_j \right|^\alpha$, which is equivalent to solve the equation $f'(\mu) = 0$ with $f(\mu) = \sum_{u \in V} p_{u,j}^m \left| y_u - \mu \right|^\alpha$.

We have $f'(\mu) = -\alpha \sum_{u \in V} (p_{u,j})^m \operatorname{sgn}(y_u - \mu) \left| y_u - \mu \right|^{\alpha - 1}$ and $f''(\mu) = \alpha(\alpha - 1) \sum_{u \in V} (p_{u,j})^m \left| y_u - \mu \right|^{\alpha - 2}$.

One can easily show that the Newton-Raphson algorithm does not converge when $\alpha < 2$. Consequently, the minimization step of GFCM with fixed norm parameter works as following:

1. Let $p_{u,i}^{(0)}$ and $\mu_i^{(0)}$ be initial values;
2. If $\alpha > 2$ compute $\mu_j^{(q+1)}$ by the Newton-Raphson algorithm:
   a) Let $\mu_j^{(q+1,0)}$ be an initial value;
   b) Do: $\mu_j^{(q+1,k+1)} = \mu_j^{(q+1,k)} + \frac{\sum_{u \in V} (p_{u,j}^{(q)})^m \operatorname{sgn}\left(y_u - \mu_j^{(q+1,k)}\right) \left| y_u - \mu_j^{(q+1,k)} \right|^{\alpha-1}}{(\alpha-1) \sum_{u \in V} (p_{u,j}^{(q)})^m \left| y_u - \mu_j^{(q+1,k)} \right|^{\alpha-2}}$ until convergence. $\mu_j^{(q+1)}$ is the limit of this sequence.
3. If $\alpha < 2$, compute $\mu_j^{(q+1)}$ by Gradient descent algorithm:
   a) Let $\mu_j^{(q+1,0)}$ be an initial value;
   b) Do: $\mu_j^{(q+1,k+1)} = \mu_j^{(q+1,k)} + \varepsilon \sum_{u \in V} (p_{u,j}^{(q)})^m \operatorname{sgn}\left(y_u - \mu_j^{(q+1,k)}\right) \left| y_u - \mu_j^{(q+1,k)} \right|^{\alpha-1}$ where ε is a fixed temporal step. $\mu_j^{(q+1)}$ is the limit of this sequence.
4. Compute $p_{u,j}^{(q+1)} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left| y_u - \mu_j^{(q+1)} \right|}{\left| y_u - \mu_k^{(q+1)} \right|} \right)^{\frac{\alpha}{m-1}}}$

D. Eulerian functions

The Eulerian function is defined as an integral for any complex number which real part is strictly positive as:

$\Gamma(z) = \int_0^{+\infty} t^{z-1} \exp(-t) dt$.

For any strictly positive integer, we have $\Gamma(n) = (n-1)!$ and for any complex z such that $\mathrm{Re}(z) > 0$, we have $\Gamma(z) = \frac{\Gamma(z+1)}{z}$. Consequently, Γ admits a meromorphic extension to the complex plane whose singularities are negative or null integers. The infinite product of Γ is given by:

$\Gamma(z) = \frac{e^{-\gamma z}}{z} \times \prod_{n=1}^{\infty} \frac{e^{\frac{z}{n}}}{1+\frac{z}{n}}$, where $\gamma = \lim_{n \to +\infty} \left[\sum_{k=1}^{n} \frac{1}{k} - \log(n) \right]$.

We define the di-gamma function as $\psi = \frac{\Gamma'}{\Gamma}$. It is also a meromorphic function which Laurent development is given by:

$\psi(z) = -\gamma - \frac{1}{z} + \sum_{n=1}^{+\infty} \frac{z}{n(z+n)}$. The consecutive derivatives of ψ are given by the Laurent developments:

$\psi^{(k)}(z) = (-1)^{k+1} k! \times \sum_{n=0}^{+\infty} \frac{1}{(z+n)^{k+1}}$.

E. Fisher information matrix, Kullback information divergence and related results

Let $\Lambda = \{ y \to p(y|\theta) : \theta \in \Theta \}$ be a smooth manifold of statistical distribution parameterized by an open set $\Theta \subset \Re^k$, the Fisher information matrix for the value θ of the parameter is given by:

$I_{i,j}(\theta) = -E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(Y|\theta) \right]$,

which is the negative of the mean of the Hessian of the log-likelihood. Under good conditions (reversibility of integration and derivation), this matrix is strictly positive and symmetric.

Each statistical distribution $y \to p(y|\theta)$ lies on an embedding set of infinite dimension; indeed to represent the entire graph of such a function, we need an infinite number of values for $_y$. However, it is parameterized by a finite number of real numbers; consequently, $_\Lambda$ has an intrinsic dimension equal to the number k of real parameters. A Riemannian manifold is provided with an infinitesimal distance which, in our case, is given by the Fisher information matrix. Without giving all the details regarding the differential geometry, one can say colloquially that the distance between "close" distributions $y \to p(y|\theta)$ and $y \to p(y|\theta + d\theta)$ is given by:

$dl = \sqrt{(d\theta)^* I(\theta) d\theta}$. Let $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$ be two values of the parameter, the length of the curve $t \in [t_1, t_2] \to \theta(t)$ where $\theta(t_1) = \theta_1$ and $\theta(t_2) = \theta_2$ in the space of distributions is given by:

$$L(\theta) = \int_{t_1}^{t_2} \sqrt{(\theta'(t))^* I(\theta(t)) \theta'(t)} dt,$$

where $t \to \theta'(t)$ is the derivative of $t \to \theta(t)$ along $_t$. The distance between the distributions $y \to p(y|\theta_1)$ and $y \to p(y|\theta_2)$ is the length of the smallest curve $t \to \theta(t)$.

The Kullback divergence between two probability densities p (target probability) and q (instrumental probability) is defined as:

$K(p:q) = \int_{\Re} \log\left(\frac{p(y)}{q(y)}\right) p(y) dy.$ The Kullback divergence is not a metric and $K(p:q) \neq K(q:p)$. However, if p and q are in the same parametrical set, denoting $K(\theta_2 : \theta_1)$ the Kullback divergence for $q = p(\ |\theta_1)$ and $p = p(\ |\theta_2)$, the Kullback divergence satisfies the asymptotic equation:

847 $K(\theta + d\theta : \theta) = (d\theta)^* I(\theta) d\theta + \circ(\|\theta\|^2)$,

848 when $d\theta$ tends to 0.